DOCUMENT RESUME

ED 460 965 TM 030 061

| | |
|---|---|
| AUTHOR | Klein, Stephen P.; Hamilton, Laura |
| TITLE | Large-Scale Testing: Current Practices and New Directions. |
| INSTITUTION | Rand Corp., Santa Monica, CA. |
| PUB DATE | 1999-04-00 |
| NOTE | 49p.; An earlier version of this paper was prepared for the Consortium on Renewing Education, Peabody Center for Education Policy, Peabody College, Vanderbilt University, under a grant from the Ball Foundation. |
| AVAILABLE FROM | Rand Education, 1700 Main St., P.O. Box 2138, Santa Monica, CA 90407-2138. Tel: 310-393-0411. |
| PUB TYPE | Reports - Evaluative (142) |
| EDRS PRICE | EDRS Price MF01 Plus Postage. PC Not Available from EDRS. |
| DESCRIPTORS | Academic Achievement; *Achievement Tests; Adaptive Testing; *Computer Assisted Testing; Criteria; *Educational Testing; Elementary Secondary Education; Internet; *National Competency Tests; Standardized Tests; Standards; *Test Use |
| IDENTIFIERS | *Large Scale Assessment; Large Scale Programs; National Assessment of Educational Progress |

ABSTRACT

This paper reviews the salient characteristics of the current inventory of large-scale achievement tests, analyzes two recent proposals for creating a national testing program, and describes a new approach to statewide and nationwide testing that RAND is examining. Section 2 discusses the criteria typically used for evaluating large-scale testing programs, and section 3 considers current tests in light of the evaluation criteria. Section 4 describes the two methods proposed for obtaining valid measures of individual student achievement that can be used to monitor student progress relative to national standards: (1) linking existing state and local tests to the National Assessment of Educational Progress; and (2) the Voluntary National Test proposed by President Clinton. Section 5 presents a new and promising method, one based on using computerized adaptive tests delivered over the Internet. Section 6 summarizes the advantages of the proposed approach and notes issues that need to be resolved regarding its implementation. (Contains 55 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 460 965

# Large-Scale Testing

## *Current Practices and New Directions*

STEPHEN P. KLEIN
LAURA HAMILTON

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

J. Ryan

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM030061

RAND
EDUCATION

BEST COPY AVAILABLE

RAND issue papers explore topics of interest to the policy-making community. Although issue papers are formally reviewed, authors have substantial latitude to express provocative views without doing full justice to other perspectives. The views and conclusions expressed in issue papers are those of the authors and do not necessarily represent those of RAND or its research sponsors.

© Copyright RAND 1999. All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from RAND.

RAND is a nonprofit institution that helps improve policy and decisionmaking through research and analysis. RAND® is a registered trademark. RAND's publications do not necessarily reflect the opinions or policies of its research sponsors. RAND URL: www.rand.org

April 1999

# Large-Scale Testing

## *Current Practices and New Directions*

By
Stephen P. Klein
and
Laura Hamilton

RAND
EDUCATION

Building on more than 25 years of research and evaluation work, RAND Education has as its mission the improvement of educational policy and practice in formal and informal settings from early childhood on.

## Acknowledgments

An earlier version of this paper was prepared for the Consortium on Renewing Education (CORE), Peabody Center for Education Policy, Peabody College, Vanderbilt University, under a grant from the Ball Foundation. We are grateful to Brian Stecher, Mark Berends, Laura Zakaras, and Jeri O'Donnell for their numerous suggestions that improved this paper.

# 1. Introduction

Policymakers and educators have identified a number of approaches to reforming the nation's public schools. These range from voucher systems and teacher training initiatives to standards-based accountability systems. The ultimate test of a reform's success is the degree to which student achievement is improved. Consequently, student assessment is at the heart of virtually all educational reforms.

Many policymakers want to use tests to assess how well schools fare when compared to schools in the same district or state or in the nation. This information could be used to motivate school personnel, to reward successful efforts, and to indicate where additional resources or changes in practices are most needed. In fact, one of the primary benefits of such an accountability system, according to its advocates, is that it would serve as an impetus for reform by mobilizing teachers, parents, and members of the local community to support educational improvement (Smith, Stevenson, & Li, 1998). Of course, there are risks that test scores also could be used for less benign purposes, such as penalizing low-performing schools.

In his 1997 State of the Union address, President Clinton argued the merits of developing a national test for fourth grade reading and eighth grade math. This set of examinations would be the first in this country to offer individual student scores on a common test for all students, and would provide students, parents, and teachers with information regarding how well students are performing in relation to their peers across the country. Presumably, this information could help parents and teachers identify stu-

dents who are falling behind in important skills so that the necessary help could be provided (Clinton, 1997). This proposal was intended to "help ensure that all of America's children have the opportunity to achieve academic success in reading and mathematics" (Smith, Stevenson, & Li, 1998, p. 42).

All of these proposed uses assume that tests provide valid and cost-effective indicators of student proficiency. This paper reviews the salient characteristics of the current inventory of tests, analyzes two recent proposals for creating a national testing program, including Clinton's Voluntary National Test (VNT), and describes a new approach to both statewide and nationwide testing that RAND is currently examining.

Section II of this paper discusses the criteria typically used for evaluating large-scale testing programs. Those familiar with such criteria should skip that section and turn directly to Section III, where tests presently used for large-scale assessment programs are discussed in light of the evaluation criteria. Section IV describes two methods proposed for obtaining valid measures of individual student achievement that can be used to monitor pupil progress relative to national standards: (1) linking existing state and local tests to the National Assessment of Educational Progress (NAEP), and (2) the VNT. Section V presents a promising new, alternative method, one that is based on using computerized adaptive tests (CATs) delivered over the Internet. We believe that this approach may offer a more-valid and reliable way to measure what students learn. Finally, Section VI summarizes the advantages of our proposed approach and notes issues that need to be resolved regarding its implementation.

## II. Criteria for Evaluating Large-Scale Achievement Tests

In evaluating the quality of an assessment instrument, one must first understand the overall purposes educators and policymakers have in mind for the test. Criteria used to judge the quality of a test for determining whether students graduate will be somewhat different from those used to judge the quality of a test for assessing the efficacy of various educational programs. And some tests are used for more than one purpose. In his VNT proposal, for example, Clinton (1997) stated that "good tests will show us who needs help, what changes in teaching to make, and which schools need to improve. They can help us to end social promotion." Scores from the VNT would, according to Clinton, be used to identify students as well as schools that are experiencing difficulties, and would have high stakes for students in that poor performance could result in failure to advance to the next grade level. Test results also can be incorporated into existing local and state accountability systems that provide rewards and sanctions to schools based on student performance.

Additionally, if a test is used to make important decisions about individual students—such as whether they should be promoted or graduated—then students should be offered several opportunities to pass truly comparable versions of the test (National Research Council, Committee on Appropriate Test Use, 1998). Students' test results also should be reported shortly after test administration.

Table 1 lists and briefly defines the major criteria typically considered in judging the technical quality of a test. More extensive discussions of these criteria

**Validity**: Degree to which a test measures what it was intended to measure.

**Reliability**: Consistency and dependability of assessment results.

**Equity**: Fairness, lack of bias toward any particular group of examinees.

**Costs and feasibility**: Monetary and other (e.g., opportunity) costs of administering a test.

**Test security and standardization**: Degree to which inappropriate exposure to test questions is prevented and consistent administration conditions are maintained.

**Mechanisms for calibration across administrations**: Procedure for ensuring that scores obtained on different administrations of the test (including those using alternate forms) are comparable with one another; e.g., there is adjustment for any differences in difficulty across different versions (forms) of the test.

**Use of relative or fixed standards**: Way in which scores are interpreted—compared with performance of other examinees or with a prespecified standard such as a cut score.

**Alignment with curriculum**: For tests intended to measure outcomes related to a particular instructional program, the degree to which items reflect what was actually taught.

**Consequences**: Inferences and actions that result from the use of the test scores.

**Timeliness of score reporting**: Length of time between test administration and score reporting.

**Motivation**: Degree to which examinees are motivated to do their best.

**Representativeness of examinee samples**: Whether the students taking the test are similar to the entire population about which inferences are being made.

**Public trust and acceptability**

**Legal defensibility**

Table 1. Criteria for Evaluating Large-Scale Assessments

can be found in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985). Note that although our focus here (and thus the title of the table) is confined to tests used in large-scale assessment programs, many of these criteria are also relevant for other kinds of tests, such as those created by teachers for the purpose of assigning grades. Our list is not exhaustive, but it does include features that must be examined when evaluating testing programs.

The methods used to achieve high levels of validity, reliability, and equity (or fairness) affect the number of questions asked; range of knowledge, skills, and abilities assessed; question format (e.g., multiple-choice versus other response formats, such as essay); administration and scoring procedures; costs; etc. Consideration of these factors often requires difficult trade-offs. For instance, longer tests tend to produce more-reliable scores and can sample a wider range of content, two important considerations given that curricular differences among schools can significantly affect what students have had the opportunity to learn. However, longer tests generally require more classroom testing time. Longer tests and those on which students must produce their own answers also tend to cost more than shorter ones and those on which students merely select their preferred answers from the choices offered (Stecher & Klein, 1997; Wainer & Thissen, 1993).

*Schools or educational programs can be fairly compared with one another only if the students tested are representative of their entire populations of students.*

Some testing programs do not require scores for individual pupils. For example, different students may take different sections of a larger test if the school, district, or state simply wants to monitor how well its students are performing as a group. One major advantage of this technique, called *matrix sampling*, is that it allows the testing program to assess a much broader range of knowledge, skills, and abilities than is feasible if individual student scores are required. One requirement for tests designed to provide indicators of performance at the group level, however, is that the sample of students be representative: Schools or educational programs can be fairly compared with one another only if the students tested are representative of their entire populations of students.

A critical component of an assessment program that uses different sets of items across administrations is a mechanism for calibrating, or what is technically known as *equating*, scores obtained on one administration to those obtained on other administrations. Such a mechanism is essential if the scores are to be used to monitor changes in student or school performance over time. This is especially the case in large-scale assessment programs, where test security, and thereby test validity, requires that questions asked on one administration are generally different from those asked on other administrations. These differences in test content may make questions asked on one occasion, as a group, somewhat harder than those asked on another. Equating is designed to adjust the scores to correct for these differences in difficulty.

Test scores can be interpreted in terms of relative or fixed standards. Relative, or *norm-referenced*, judgments involve examining how well a student performs

in comparison to other students. For example, test scores are often used to gauge how well the students at a school perform relative to students in some national sample of students or to students at other schools in the same district or state. In contrast, fixed, or *criterion-referenced*, judgments involve examining the degree to which students have mastered a specific body of knowledge, skills, and abilities. These types of comparisons are used in standards-based testing programs that seek to provide information on numbers of students who have achieved a preset standard of performance. Both norm- and criterion-referenced interpretations assume that students take the same or truly comparable tests, that these tests are administered under the same conditions (e.g., time limits), and in the case of open-ended exams, that the standards used to evaluate answer quality are the same for all students.

Additional criteria for evaluating tests include the degree to which a test is aligned with the curriculum (which is particularly important for tests intended to measure effects of specific educational programs), consequences of test use, and timeliness of score reporting (an important consideration for scores intended to inform decisions about individual pupils). The level of student motivation also must be considered: When a testing situation has low stakes for students, they may not put forth their best effort.

The extent to which a sample of test-takers is representative of a larger population is an important criterion in evaluating tests used to make inferences about the performance of groups, such as students in a particular district or state. Finally, public acceptability and legal defensibility are necessary considerations in almost any large-scale assessment program.

## III. The Utility of Existing Large-Scale Tests

The United States currently has no shortage of large-scale, standardized achievement tests. However, the existing inventory of tests does not provide comparable individual scores for the entire population of students, and there are good reasons to believe that even with extensive modifications, no one test could. In this section, we discuss the utility of the most widely used national and international achievement tests. We then discuss state tests, including some of the commercially available standardized achievement tests. Our discussion does not explicitly apply each criterion described in Section II to every test. Instead, the focus is on each test's major strengths and limitations in light of those criteria.

### National and International Achievement Tests

The National Assessment of Educational Progress (NAEP) was established to monitor state and national achievement trends. The "state" portion of this program, which provides scores for individual states, has administered tests in one or two subjects every two years since 1990. As Table 2 shows, the grade levels and subjects tested vary on a yearly basis. For example, in 1996, State NAEP administered eighth-grade math and science tests and fourth-grade math tests; in 1998, State NAEP administered eighth-grade reading and writing tests and fourth-grade reading tests. The national portion of NAEP provides national scores and enables trend analysis. It also frequently administers tests in subjects other than those covered by State NAEP, such as history and social studies.

| | 1990 | 1992 | 1994 | 1996 | 1998 |
|---|---|---|---|---|---|
| Reading | | 4 | 4 | | 4&8 |
| Math | 8 | 4 & 8 | | 4 & 8 | |
| Writing | | | | | 8 |
| Science | | | | 8 | |

*Source: U.S. Department of Education (1997).*

**Table 2. Grade Levels and Subjects Tested in State NAEP, 1990–1998**

Only State NAEP, however, provides the data needed to make comparisons among states.

Several features of the State and National NAEP help to ensure the validity of their test results. Samples of students are carefully selected to be representative of the state's (or nation's) students, items are matrix sampled (i.e., different students answer different sets of questions) so that a wide range of content and skills can be assessed, and administration conditions are carefully controlled to ensure test security and standardization across schools. For example, specially trained NAEP staff and consultants, rather than classroom teachers, administer the tests. NAEP also collects information on family background, teaching practices, and other variables that aid in understanding results, though these measures do not capture all of the background information needed to explain test score trends. The primary limitation of NAEP, and the one that was the impetus for the VNT proposal, is that NAEP data cannot be used to monitor the progress of individual students, schools, or

BEST COPY AVAILABLE

even districts within a state, because NAEP tests are only administered to a random sample of the students in each state that agrees to participate.

Some tests have been devised specifically for making international comparisons. Two recent examples are the International Association for Evaluation of Educational Achievement Reading Literacy Study (Elley, 1993) and the Third International Math and Science Study (TIMSS; see Mullis et al., 1998). These efforts involve testing representative samples of students from participating countries, including the United States. Test results are used primarily for evaluating the relative standing of U.S. students compared with their counterparts in other countries. These testing programs are similar to NAEP in that scores are not reported for individuals, schools, districts, or states, which means that no real consequences are associated with good or poor performance for anyone involved in taking these tests.

Other large-scale standardized tests have been developed primarily for research purposes. The National Center for Education Statistics (NCES) has carried out a series of longitudinal studies, the most recent being the National Education Longitudinal Study of 1988 (NELS:88). These studies involve measuring student achievement in several subjects. In NELS:88, for example, tests were administered in four subjects to samples of students at grades 8, 10, and 12 (Owings, 1995). In addition, questionnaire data were collected from students, teachers, principals, and parents. The resulting database enables researchers to explore a wide variety of issues, such as relationships between test scores and various student characteristics and experiences. The longitudinal design of these studies allows researchers to track stu-

dent achievement over time. However, scores are not given to students, parents, or schools.

All of the tests discussed thus far—NAEP, TIMSS, and NELS—are administered to nationally representative samples of students under highly controlled, standardized conditions. But again, the scores on these tests are not used to make decisions about individual students. The question of motivation thus arises, since students may not be motivated to perform their best when there are no personal stakes involved, and, consequently, their scores may not accurately reflect their true capabilities (O'Neil et al., 1997). The next type of tests discussed are those used for selection purposes, which means they do involve high stakes for students.

At the national level, the most notable examples of tests that involve high stakes are the college entrance examinations, which include the Scholastic Assessment Test (SAT) I (general) and II (subject tests) and the American College Testing (ACT) program tests. Students electing to take these tests are generally motivated to perform well because their scores influence their options for post-secondary education. Aggregate scores on these exams are sometimes used to monitor the college-bound population and to make comparisons among states, districts, and schools. However, whereas the samples of students taking NAEP and the international tests described above are representative, the samples of students taking the SAT and ACT exams are neither locally nor nationally representative. Moreover, the degree of selection and the factors governing selection are likely to vary across regions. In some states, for example, nearly all college-bound students take the SAT. In other states, whose higher education systems use the ACT, the

SAT-taking population may be largely limited to students planning to attend out-of-state schools. Consequently, SAT and ACT scores are not appropriate for measuring achievement trends or comparing states or districts.

There are similar "selection bias" concerns with monitoring scores on the Armed Forces Qualification Test (AFQT), which is taken by men and women interested in enlisting in one of the U.S. military services.

*...none of the large-scale national achievement tests currently in use can be employed to monitor individual student progress or to evaluate the effectiveness of particular schools, districts, or educational programs.*

In summary, then, we can say that despite their variety, none of the large-scale national achievement tests currently in use can be employed to monitor individual student progress or to evaluate the effectiveness of particular schools, districts, or educational programs. The state tests described below are usually intended to serve one or more of these purposes.

## State Tests

Almost every state has or is developing a set of statewide tests (American Federation of Teachers, 1997; *Education Week*, 1999). These assessments vary in scope and purpose: Some states produce individual student scores and use them to determine whether students may advance to a new grade level or graduate from high school; other states use their assessment results to distribute rewards and sanctions to schools based on changes in performance. Test scores also have been used to evaluate teacher effectiveness and to monitor student progress toward state or district goals.

The format of these assessments also varies. Some states use a commercially available, standardized multiple-choice test, whereas others develop their own tests or assessment methods, including portfolios or other open-ended response formats. A test's format often reflects a desire to evaluate the success of a reform effort. The Kentucky Instructional Results Information System (KIRIS), for example, was designed to monitor the effects of the Kentucky Education Reform Act. This objective led KIRIS to emphasize the kinds of open-ended tasks that were presumed to reflect reform-related instruction. The portfolios used in some states and districts also resulted from efforts to align assessment instruments with instruction and to evaluate students in an "authentic" way. Unfortunately, concerns about authenticity have often taken undue precedence over the technical quality of the assessments, resulting in tests that fail to provide valid or useful information (Klein, McCaffrey, Stecher, & Koretz, 1995; Koretz & Barron, 1998; Koretz, Stecher, Klein, & McCaffrey, 1994).

States sometimes change the nature of their tests from one year to the next. Kentucky eliminated multiple-choice items from KIRIS but then reintroduced them after an expert blue-ribbon panel found that such items were needed to ensure content coverage, score reliability, and stability of proficiency standards over time (Hambleton et al., 1995). California is now using a commercial multiple-choice test after pioneering the controversial California Learning Assessment System (CLAS) tests, which included both open-ended and multiple-choice items. Such changes in the measures used often reflect responses to political pressure, changes in curriculum, financial factors,

or technical concerns. Whatever the reason, such changes make it difficult—if not impossible—to track achievement from one year to the next. However, as discussed below, when states use the same test for several years in a row, they are likely to encounter problems that threaten or even completely undermine the validity of the scores obtained. This situation led the National Research Council's Committee on Appropriate Test Use (1998) to recommend that a different test be used each year and that these tests have comparable content coverage and be equated to each other.

Despite the considerable resources devoted to state testing programs each year, most state tests have several serious shortcomings that prevent them from providing trustworthy data on student achievement. The most pervasive of these problems are compromised test security, narrow focus of measurement, and low student motivation. Several state programs suffer from most or all of these problems; others have taken steps to try to eliminate or at least mitigate them.

*...most state tests have several serious shortcomings that prevent them from providing trustworthy data on student achievement.*

### *Compromised Security*

Because of the expense involved in hiring external test administrators, most states have classroom teachers administer the state tests to their own students. Administration typically involves teachers following a set of clearly specified steps and reading instructions to their students from a "script." These carefully designed procedures are intended to ensure that all students receive the same instructions across

classrooms and schools. This standardization (including the amount of time students are given to respond) is essential if the scores of students from different schools and classrooms are to be validly interpreted.

Although it is reasonable to expect that most teachers will carry out this activity competently, test security and standardization may be compromised when teachers administer tests to students in their own schools. How teachers adhere to time limits and respond to student questions are just two of the many factors that have been found to vary across testing sites (see, e.g., Koretz, Mitchell, Barron, & Keith, 1996), and in some cases outright cheating has been reported (Stecklow, 1997). This variability in administration conditions compromises score comparability, and the resulting lack of public trust in the scores defeats the purpose of the assessment. Serious problems can also occur when test administration involves setting up or demonstrating equipment, or when students are given complex instructions. Another threat to consistency in administration conditions occurs when products (such as book reports) that are created in class or at home are used as part of the testing program. For example, in creating such products, some students may receive more teacher or parent assistance than others do, rendering comparisons among students inappropriate (Gearhart, Herman, Baker, & Whittaker, 1994).

We analyzed test data from several districts and states that have large-scale testing programs. We also gave other tests at these sites, using RAND-trained administrators, so that we could compare test scores under two quite different administration conditions. The results of this field experiment revealed that scores on the externally administered tests consistent-

ly showed the expected relationships among themselves and with various student characteristics, such as socioeconomic status. This was true regardless of format (e.g., multiple choice or open response). In contrast, several of the locally administered tests showed markedly different relationships with student characteristics and unexpectedly low correlations with the external measures.[1] This evidence suggests that the administration of the local assessments was compromised and that scores on statewide or district tests may not be trustworthy (see also Cannell, 1987 and 1989). The result of such breaches in test security and administration conditions is that student scores often reflect no more than how well those students can answer the particular questions asked on the test. That is, the scores cannot be used to make generalizations about how students are doing in the subject matter area (e.g., mathematics) being assessed (Mehrens & Kaminski, 1989).

*This problem of "teaching to the test" is exacerbated when states and districts ask exactly the same questions year after year.*

In short, teaching students how to answer specific questions that appear on a test (or ones very similar to those questions) defeats the very purpose of the test. This problem of "teaching to the test" is exacerbated when states and districts ask exactly the same questions year after year. Linn, Graue, and Sanders (1990) showed that when a state or district switches from one test form to another, there is an initial sharp drop in scores followed by a gradual increase over the following few years. As the test becomes more familiar, scores increase. In Kentucky,

[1]These findings have not been published, but tables showing these trends are available from the authors.

Hambleton et al. (1995) found that the improvement in performance on items that had also been administered the previous year was artificially high, and suggested that coaching on these items might have occurred (see also Koretz & Barron, 1998). Even well-meaning teachers are likely to incorporate their knowledge of test items into their instruction, especially if high stakes are attached to results for them or their students (Smith, 1991). It is therefore critical that the questions be changed each time the test is administered (as per the methods commonly used on high-stakes college admissions tests and licensing exams for doctors and lawyers). To track student progress, however, the different questions must be statistically equated so that test scores are on a common scale (Angoff, 1971; Kolen & Brennan, 1995).

### *Narrow Focus of Measurement*

One concern often voiced about the use of standardized tests is the limited content coverage of any single test (Jones, 1997; Stake, 1995). Because of the constraint of limited testing time, it is difficult to sample adequately from each content area, even when tests are designed to balance item content across a specified set of topics. NAEP solves this problem through matrix sampling; i.e., it administers a large number of items from each content area but requires any given student to answer only a small portion of the total set of items. This strategy permits more accurate inferences about typical student achievement than can be obtained when the same set of questions is administered to every student.

The multiple-choice format also may limit the kinds of skills that can be assessed. Many states have

supplemented or replaced multiple-choice items with open-ended items, such as essays, hands-on experiments, or portfolios. Although combining several formats can enhance coverage, the addition of open-ended items causes other problems. They take longer to administer than do multiple-choice items, so fewer questions can be asked in a given amount of testing time. Furthermore, they do not necessarily measure the kinds of skills that their developers intended or that some advocates claim they do (Hamilton, Nussbaum, & Snow, 1997). And the substantial cost of scoring such items also introduces trade-offs between finances and test reliability (Stecher & Klein, 1997; Wainer & Thissen, 1993). The result is that attempts to increase breadth by using multiple formats are often made at some expense and with little assurance that the measurement goals have been met.

### *Low Student Motivation*

Some states feel they provide a form of motivation when they create high stakes for schools and teachers by tying financial rewards and sanctions to student performance. It is often argued that such "accountability" systems ensure that all students receive high-quality instruction, and there is some evidence that teachers do alter their practices in desirable ways in response to external tests (Koretz, Stecher, Klein, & McCaffrey, 1994; Stecher, Barron, Kaganoff, & Goodwin, 1998). But in many cases, these systems lead only to artificial inflation of scores with no real evidence of improved learning (Koretz & Barron, 1998).

As for how stakes are involved directly for students, the answer varies from state to state. Some states make graduation, placement in an advanced

track, or promotion to the next grade contingent upon a student's achieving a particular score, whereas other states attach virtually no consequences to student scores. Differences of these types contribute to variation in the degree to which students are likely to be motivated to perform well, which has been shown to affect scores. Wainer (1993) found that when students perceive tests as having no direct consequences, they may not put forth as much effort as they would when the stakes are higher. Wolf and Smith (1995) found that the performance of college students was roughly one-quarter of a standard deviation higher when a test counted for their grade than when it did not. O'Neil et al. (1997) also found that paying eighth graders for correct answers on a NAEP examination resulted in higher scores (but this did not happen with twelfth graders). Interpretations of most achievement test results rely on the assumption that students have given their best efforts. The results from these studies illustrate that this assumption may not always be correct (see also Brown & Walberg, 1993).

Differences in administration conditions, item content and format, student motivation, and previous exposure to a test's questions make it impossible to compare results across states reliably. Nor can one compare results within states over time, or a state's results to any set of national standards. And the same is true for comparisons at the school and district levels. Furthermore, the test-administration conditions in many states preclude valid comparisons among individual students or even schools or districts. Ways proposed for

*Differences in administration conditions, item content and format, student motivation, and previous exposure to a test's questions make it impossible to compare results across states reliably.*

addressing these problems are discussed in the next two sections.

## IV. Two Proposed Approaches to a National Measure of Achievement

Policymakers and the education community have recently considered two major approaches to providing a common, national measure of achievement. One strategy is to allow states to continue administering their own tests, and then to link the results of those tests to NAEP. The second approach, advocated by President Clinton in his 1997 State of the Union address, is to create a national test. We discuss both of these options.

### Linking State Tests to NAEP

The term *linking*, when used in a measurement context, generally refers to a process whereby results on one assessment are made comparable to those on another. There are multiple ways to perform a linking analysis, each of which carries certain assumptions and technical requirements (Kolen & Brennan, 1995). Linn (1993), for example, describes five forms of linking and compares them in terms of statistical rigor and the types of comparisons for which they can be used. The product of a linking analysis is generally a set of equations or a table that allows the user to convert a student's score on one test (such as an exam administered by a state) to a score on another test that the student did not take (such as NAEP).

As discussed above, the significant differences among state testing programs preclude making any

direct comparisons of results across states. However, NAEP scores provide a potentially useful benchmark for putting the results of different tests on a common scale. This approach is analogous to measuring the value of the currency of different countries by comparing the worth of each country's currency to that of the U.S. dollar. In short, the goal is to transform a student's score on the state test into a score on a NAEP-based scale so that parents, schools, and teachers will be able to see how well their students are doing relative to a national set of standards. In a recent survey of state assessment and curriculum directors and other users of NAEP results, participants expressed support for this type of linking (Levine, Rathbun, Selden, & Davis, 1998).

Linking is often used in other testing situations. Each time the SAT is administered, for example, new test forms are created and the resulting scores are placed on a scale that allows them to be compared with scores obtained on previous administrations. The process of rendering alternate forms of the same test comparable so that their scores can be used interchangeably is typically called *equating*, but can be thought of more generally as a special case of linking. Because equating requires that the tests measure the same constructs in an equally reliable way, the linking of a state test to NAEP would not be considered true equating. Linking can still be done, however, and can, at least in theory, yield useful results, albeit weaker ones than are obtained through a strict equating procedure.

Several studies have been conducted to investigate the feasibility of linking state assessment results to NAEP. The results of Ercikan (1997) and of Linn and Kiplinger (1995) point to the instability of the link

between state tests and NAEP: The NAEP scores derived from the linking differ substantially from actual NAEP scores, particularly at the extreme ends of the score distribution. Waltman (1997) found that the use of linking functions to classify students into performance categories tends to result in low levels of agreement between the state test and NAEP. Problems have also been encountered in attempts to link international assessments to NAEP (see, e.g., Beaton & Gonzales, 1993; Pashley & Phillips, 1993).

Most recently, the National Research Council convened a panel of experts to investigate the feasibility of linking. The panel concluded that linking is not a viable option because of differences in content, format, difficulty, measurement precision, and administration conditions (particularly test uses and consequences) across states (National Research Council, Committee on Equivalency and Linkage of Educational Tests, 1998).

Plausible explanations exist for why efforts to link tests have not been successful. We discuss several of these explanations here, including differences in content and format, variations in representativeness of student samples, variations in administration conditions and student motivation, reuse of test forms at the state level, and reporting time. Most of these are also discussed in the report prepared by the National Research Council's Committee on Equivalency and Linkage of Educational Tests (1998).[2]

One of the most obvious problems with linking is that the majority of state tests differ from NAEP in content, format, or both. A study by Bond and Jaeger

[2]The NRC report was published a few months after the initial draft of this paper was circulated.

(1993), in which content experts classified items on three standardized tests into content categories, revealed wide disparities in the balance of items across categories in NAEP versus the standardized tests. However, a study by Linn and Kiplinger (1995) revealed that even when the NAEP scale used is aligned with the content of a state test, the resulting linking errors are of the same magnitude as those resulting when content is not considered. Thus, "alignment" alone does not explain why linking failed. Before defensible links can be constructed between NAEP and existing state tests, more research will be needed to understand how content and format differences affect linking.

The representativeness of student samples is another factor that may affect the quality of linking. For NAEP, schools are sampled to be roughly representative of the state's population of students, but this sample may not represent the population of students who actually take the state test. For example, states differ in their policies regarding the inclusion in the testing program of students who have limited English proficiency. Schools also differ in the efforts they make (such as offering make-up testing) to ensure full participation among their students. The especially high absentee rates at some schools on the day the test is administered raise concerns about the validity of school-level data.

Differences in administration conditions and student motivation may also affect linking. Bloxom, Pashley, Nicewander, and Yan (1995), for example, linked NAEP scores to scores on the Armed Services Vocational Aptitude Battery (ASVAB). Examination of the score distributions on the two tests suggested that examinees were more motivated to perform well

on the ASVAB than on NAEP, which is consistent with the fact that only the ASVAB scores had consequences for these examinees. It is likely that similar problems would be encountered when scores on other high-stakes tests are linked to NAEP scores.

State results are influenced by the number of times a particular test form has been used. Unlike NAEP, many states administer the same form of a test for several years in a row. As discussed above, Linn, Graue, and Sanders (1990) showed that scores increase as a form is reused, particularly during the first few years. Thus, repeated use of some state tests is another potential source of error.

Finally, an additional administrative concern with the plan to link tests is the inevitable significant lag between test administration and reporting. Test papers must be collected, assembled, and prepared for scoring; scores must be assigned, verified, and converted to a usable data format; and reports must be generated. Because of the scope and complexity of NAEP, its results can take over a year to be released, and it is not given every year. Consequently, the linking of state results to NAEP would not occur until well after state results have been released, which could render the linking almost meaningless, at least from the public's perspective.

*...major overhauls of most existing state assessment programs, as well as tremendous expansion of NAEP, would be required to make the linking of state tests to NAEP a feasible and appropriate strategy.*

In summary, major overhauls of most existing state assessment programs, as well as tremendous expansion of NAEP, would be required to make the linking of state tests to NAEP a feasible and appropriate strategy. A common content

framework and set of test specifications, standardized administration conditions, and consistency of motivational conditions (and, therefore, stakes attached to scores) would be crucial. States would also need to develop new test forms each year and ensure that these forms were equated. It is extremely unlikely that states will reach the level of consensus needed to make this happen, and most states will not have the resources needed to ensure high-quality testing conditions for all students.

## A Voluntary National Test

Given the diversity and flux among state testing programs, some educators and policymakers have argued that what is needed is a single, common test administered across states. As noted earlier, President Clinton proposed a national testing program that would produce individual student scores for every fourth grader in reading and for every eighth grader in math. The modifier of "voluntary" was added to the name of this testing program after debate over the proposal began. The Department of Education has called Clinton's plan the Voluntary National Test (VNT) because states and districts would be allowed to choose whether to participate—federal law would not require participation. However, a state and district that chose to participate could require that all schools, teachers, and students in the specified grades participate.

Unlike state tests, many of which are chosen or developed to reflect state-developed standards, items on the VNT would be designed to correspond closely with NAEP. The goal of the VNT is to "provide students, parents, and teachers with meaningful

scores to compare individual student performance to widely accepted national and international standards and to identify students and schools that need extra help" (U.S. Department of Education and National Science Foundation, 1998).

One of the most common criticisms of the VNT is that it places too many decisions about what to measure, and consequently what to teach, in the hands of the federal government. Advocates of the VNT argue that it would be aligned with the NAEP content frameworks, which presumably reflect broad consensus about what students should know and be able to do. However, state and local education agencies continue to set their own objectives, and the importance of local control over standards and assessments has been emphasized in several recent reports, including some by groups that have promoted national standards (e.g., National Council on Education Standards and Testing, 1992; National Research Council and National Council of Teachers of Mathematics, 1997). Of course, a national test need not preclude the use of state assessments as well, but critics fear that it could lead to narrowing of instruction and an overemphasis on preparing students for one test.

Even if agreement could be reached on what topics to assess, a single 90-minute test would probably not allow sufficient content coverage, especially if a significant portion of the testing time is devoted to a small number of open-ended items. Unlike NAEP, which uses matrix sampling and therefore can administer many more items than a student could reasonably be asked to take in a classroom period, the proposed VNT would be a single test given to all students. Hence, its coverage could not be as comprehensive as that of NAEP. Important topics would have to be

omitted. Moreover, the VNT plan of testing only fourth graders in reading and eighth graders in math provides a very limited snapshot of student learning and neglects many important subjects and grade levels.

Another important issue raised by the VNT is test security. To provide valid measurement of student achievement, test forms must be kept secure. This is critically important given the consistent findings that high-stakes testing often leads to inappropriate coaching on the specific test items, resulting in scores that do not accurately reflect student proficiency (see, e.g., Koretz, Linn, Dunbar, & Shepard, 1991). Because a single form of the VNT would be given at different times across the nation and because the tests would be administered by local school personnel, there is a real risk that actual VNT items would be used in instruction or that other breaches of security (such as extensions of time limits and teacher assistance during the exam) would occur.

Finally, the costs associated with this type of large-scale testing are formidable. The cost of administering a commercially available, standardized multiple-choice test is typically between $4 and $6 per student in 1999 dollars. This figure includes the test booklet and answer sheets. There are additional charges for scoring and the reporting of results. The proposed VNT would also include open-ended items, which are likely to cost an additional $3 to $8 per pupil to score, primarily because scoring would have to be done by trained graders. An analysis of costs must also consider the time that teachers spend administering the test and the opportunity costs of having students spend classroom time taking it. Of course, any well-designed assessment program will have costs associated with it, but it is important to recognize

these costs and then weigh them against whatever benefits are likely to accrue from the testing program.

## V. A Promising Alternative: National Internet Computerized Adaptive Testing

*Advances in information technology and the rapidly growing presence of computers in schools provide opportunities to explore alternative modes of testing.*

Advances in information technology and the rapidly growing presence of computers in schools provide opportunities to explore alternative modes of testing. Increasingly, computers linked to the Internet are becoming integral components of curriculum. In 1994, 35 percent of public schools in the United States had access to the Internet. Three years later, this figure more than doubled, rising to 78 percent. Also in 1997, 98 percent of the schools with Internet access had at least one classroom connected, and 43 percent had five or more (National Center for Education Statistics, 1998). Elementary and secondary students currently use computers for research, document preparation, communication, learning games, drill, and other school-related activities (Collis et al., 1996; Wenglinsky, 1998).

Computers are also becoming central in the administration of many large-scale testing programs. For example, several of the tests produced by the Educational Testing Service (ETS) and the American College Testing (ACT) program are administered via computer. The Graduate Record Examination (GRE)

is probably the most familiar example. This new administration format involves more than simply displaying the paper-and-pencil versions on a computer screen. Tests such as the GRE use a computerized adaptive testing (CAT) technology in which the presentation of an item and the decision concerning when to stop testing depend on the examinee's performance on earlier items (Bunderson, Inouye, & Olsen, 1989).

Numerous studies comparing CATs with traditional paper-and-pencil tests in the same subjects have found that the two methods rank-order students in about the same way (see Mead & Drasgow, 1993, for a review), but that the computerized approach does it in far less testing time. Bunderson, Inouye, and Olson (1989) demonstrated that the number of test items needed for a given level of precision can often be reduced by more than half with the computerized approach, thereby freeing up testing time for other educational activities.

These technological trends and the advantages offered by some forms of electronic testing led us to explore the feasibility of adopting a computerized system that could achieve the objectives sought after in pursuing a national test of achievement while avoiding the drawbacks of the linking and VNT strategies. There are, of course, other approaches that could be considered, but our evaluation of the alternatives suggests that Web-based computer adaptive testing is the most promising. Specifically, we propose that large-scale testing involve multiple-choice items presented in an adaptive mode, supplemented by open-ended and other types of items, all of which would be administered in schools via the Internet.

## How It Would Work

A system for large-scale Internet testing would use "banks" of multiple-choice and other types of items in each subject to be tested. Several thousand items could be maintained in the system, along with relevant statistics (e.g., difficulty levels) on each item. Students would sit at computer terminals in their own classrooms or in computer labs at their school to take the test. Generally, the testing session would begin with a small set of items spanning a wide but age-appropriate range of difficulty. Then, depending upon how well a student does on these questions, items of greater or lesser difficulty would be administered. This process would continue until the student's proficiency level had been identified within a prespecified level of score precision. In addition to being able to assign an overall score for a subject, the system can be designed to assign scores for subareas (e.g., in math, there might be scores for such areas as computation, estimation, and problem solving).

Currently, most CATs are administered on a standalone personal computer. To be used on a large scale, such as for a district, state, or national testing program, the system would have to be delivered into all U.S. schools in an efficient, secure, and cost-effective manner. A Web-based delivery system may be the best way to do this. Such a system could be upgraded and maintained at a central location without requiring complicated installation or modification at local school sites. In addition, data could be stored centrally, facilitating the development of systemwide norms and score summaries.

## Advantages

CATs have several advantages over paper-and-pencil multiple-choice tests. As mentioned earlier, one of the major benefits is decreased testing time. Because item difficulty is tailored to each examinee's proficiency level, students do not waste time on questions that are much too easy or too difficult for them. It takes many fewer items to achieve a desired level of score precision using a CAT than using a nonadaptive multiple-choice test. This feature not only saves time, but may minimize frustration, boredom, and test anxiety.

Another potential benefit is improved test security. Because each student within a classroom takes a different test (i.e., one tailored to his or her proficiency level), there is little risk of students being exposed to test items in advance or of teachers coaching their students on specific items. CATs are particularly useful for measuring growth over time. Students take different items on different occasions, so scores are not affected by practice. However, because all items are calibrated to the same scale, growth can be measured in a straightforward way.

CATs also offer efficient and inexpensive scoring. Scoring is done on-line, eliminating the need for packaging, shipping, and processing of answer sheets. Students could be given their results immediately after completing the tests. An Internet-based system would allow all records to be stored automatically at a central location, facilitating the production of score summaries. Teachers would have results in time to do something about students who are not progressing at the expected rate, and could incorporate this information into their instruction. Teachers

could also use results for assigning grades, so that students would be motivated to do well.

Finally, computer-based testing offers the opportunity to develop new types of questions, especially those that can assess higher-level thinking skills. For example, students could observe the effects on plant growth of various amounts of water, types of fertilizer, and exposure to sunlight in order to make inferences about the relationships among these factors. Several efforts are now under way to develop innovative items such as essays that can be machine-scored, simulations of laboratory science experiments, and other types of items that require students to produce, rather than just select, answers. Many of these efforts have sought to incorporate multimedia technology to expand the range of activities in which students can engage (see Bennett, 1998, for a summary of the possibilities offered by computer technology).

*Finally, computer-based testing offers the opportunity to develop new types of questions, especially those that can assess higher-level thinking skills.*

Computerized assessments are especially appropriate for evaluating student progress in areas where computers are used frequently, such as writing. Russell and Haney (1997), for example, found that students accustomed to using computers in their classes performed better on a writing test when they could use computers rather than paper and pencil. Students using the computers wrote more and organized their compositions better. In short, the computer can accommodate many more item types than can a paper-and-pencil test, such as tasks that involve using a mouse to move objects around on the screen, draw graphics, etc. Moreover, as instruction comes to depend more heavily on technology, assessment will

need to follow suit in order to be appropriately aligned with curriculum.

## Limitations

Perhaps the greatest obstacle to the proposed system is the diversity of the technological resources across districts, schools, and classrooms. Despite the rapid entry of computers into classrooms, some schools will probably continue to lack the resources needed to accommodate our proposed system. Students from low-income families may be especially likely to attend schools that lack the necessary resources, although a number of grant programs now target technology funds at such schools. These students may also be less likely than their advantaged counterparts to have experience using computers, so their comfort level with the technology may be lower. A similar concern applies to teachers, who also vary in their levels of experience and comfort with computers. Thus, before our proposed system is implemented in a high-stakes setting, special care would have to be taken to ensure that all students and teachers not only have an opportunity to practice using the technology, but also receive the instruction that will enable them to be comfortable with it. Studies are also needed to examine the validity of the assessment for all students.

An additional limitation to most existing CAT systems is that they rely solely on multiple-choice items. However, as discussed above, an Internet-based testing program could certainly incorporate other kinds of items, and improvements in technology for administering and scoring open-ended items will increasingly make such items cost-effective.

### What Needs to Be Done?

Research is needed to explore the technical, economic, legal, and social concerns surrounding an Internet-based testing system. However, such a system appears to solve many of the serious problems associated with the assessment approaches currently being used, as well as with those that are being proposed by others for large-scale testing programs. The next steps will involve examining the issues related to creating, implementing, and maintaining an Internet-based CAT system while simultaneously building, testing, and demonstrating prototype systems.

## VI. Conclusion

The methods being proposed for reforming and governing K–12 education implicitly assume that there are valid and appropriate ways to assess student achievement. For accountability systems, evaluations of the effects of charter schools and vouchers, and court and legislative decisions regarding how much money a school must spend to provide an "adequate" education for its students—for all of these purposes and more, valid measures of student achievement are essential. Without them, there is no way to know how well students are meeting educational goals. Consequently, extensive amounts of student and teacher time plus millions of dollars are spent each year on large-scale, state and district testing programs.

Unfortunately, these expenditures of human and financial resources often fail to provide the kinds of credible data needed to draw appropriate conclusions about individual student progress or the efficacy of

educational programs and reforms. This failure occurs in part because the stakes tied to most statewide and district tests are usually for teachers, principals, and other district staff, rather than for students. For example, students' scores generally do not affect their grades, promotion, or graduation. Consequently, test-takers may not be motivated to do their best. Moreover, teachers and other school personnel may feel pressured to ensure their students receive high scores so as, for example, to avoid being responsible for students not being promoted to the next grade level. Pressures such as these have apparently contributed to widespread breaches in both test security and the standardization of test administration conditions, which together have inflated scores and undermined their validity. Scores that may indicate how well students can answer questions about a passage they have read before, rather than how well they can read, are not credible data.

There are significant differences in goals, standards, curriculum, and instructional practices across teachers, schools, districts, and states. Thus, a single test cannot be congruent ("aligned") with all of these differences. And there are other problems with these tests: The format of most of them limits the range of relevant skills and abilities that can be measured, there is often a long delay between when the test is given and when the results are provided to schools, and students experience frustration when asked questions that are much too easy or too difficult for their particular ability level.

These considerations and other factors led us to propose that most paper-and-pencil tests be replaced with computerized adaptive tests (CATs) adminis-

tered over the Internet. The major advantages of a CAT system include the following:

- Test questions can be tailored to the district's or state's educational goals.
- Tests are targeted to each student's proficiency level, which leads to spending about half the testing time per pupil to obtain the level of score reliability produced by a traditional paper-and-pencil multiple-choice test.
- A "bank" of thousands of questions in each subject matter area can be drawn on, thereby improving test security by reducing the opportunity for students to practice taking the specific questions they will be asked on the test.
- Types of questions capable of assessing important skills that cannot be measured well with more traditional question formats can be accommodated.
- The proficiencies of new students coming into a classroom can be assessed quickly and reliably.
- Students can be tested several times during the school year and the results used in calculating an individual student's grades as well as in monitoring areas in which the class as a whole needs extra help.
- Test results can be reported to teachers, parents, and students almost instantly.

The advantages of delivering CATs directly into a student's classroom or school computer laboratory via the Internet include:

- Economies of scale can be achieved because the item bank can be refreshed from a central location.

New questions can easily be inserted into existing tests for calibration and for operational use, including items that are submitted by teachers.

- Software updates are done centrally rather than locally, and the need for expensive hardware and software at the school site is eliminated.
- Records on student performance can be stored in a centralized location for analysis, reporting, and construction of norms.
- Students need not leave their schools to take tests, nor need teachers spend time packaging and shipping testing materials.

There are, of course, many issues that need to be addressed before state or national Web-based testing becomes a reality, including costs, development of large item banks, hardware and software considerations, how well teachers and others accept the transition to CATs, and legal and licensing constraints. These and other factors are critically important to the implementation of CATs, but history suggests that despite the obstacles, technological advances are implemented quickly when they fill an important need. In short, the question is not whether but when Web-based CATs will become operational on a large-scale basis. Given that some item banks have already been constructed, that pilot tests and field demonstrations of Web-based testing are already under way, and that many schools have several classrooms connected to the Internet, we suspect that the necessary infrastructure and major components of the system will be fully functioning within the next few years.

# REFERENCES


American Federation of Teachers (1997). *Making Standards Matter 1997: An Annual Fifty-State Report on Efforts to Raise Academic Standards*. Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for Educational and Psychological Testing*. Washington, DC: Author.

Angoff, W. A. (1971). "Scales, Norms, and Equivalent Scores." In R. L. Thorndike (ed.), *Educational Measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Beaton, A. E., & Gonzalez, E. J. (1993). *Comparing the NAEP Trial State Assessment Results with the IEAP International Results*. Report prepared for the National Academy of Education Panel on the NAEP Trial State Assessment. Stanford, CA: National Academy of Education.

Bennett, R. E. (1998). *Speculations on the Future of Large-Scale Educational Testing*. Princeton, NJ: Educational Testing Service.

Bloxom, B., Pashley, P. J., Nicewander, W. A., & Yan, D. (1995). "Linking to a Large-Scale Assessment: An Empirical Evaluation." *Journal of Educational and Behavioral Statistics, 20*, 1–26.

Bond, L., & Jaeger, R. M. (1993). *Judged Congruence Between Various State Assessment Tests in Mathematics and the 1990 National Assessment of Educational Progress Item Pool for Grade 8 Mathematics*. Report prepared for the National Academy of Education Panel on the NAEP Trial State Assessment. Stanford, CA: National Academy of Education.

Brown, S. M., & Walberg, H. J. (1993). "Motivational Effects on Test Scores of Elementary Students." *Journal of Educational Research, 86*, 133–136.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). "The Four Generations of Computerized Educational Measurement." In R. L. Linn (ed.), *Educational Measurement* (3rd ed., pp. 367–407). New York: Macmillan.

Cannell, J. J. (1987). *Nationally Normed Elementary Achievement Testing in America's Public Schools: How All Fifty States Are Above the National Average*. Daniels, WV: Friends for Education.

Cannell, J. J. (1989). *How Public Educators Cheat on Standardized Achievement Tests*. Albuquerque, NM: Friends for Education.

Clinton, W. J. (1997). *State of the Union Address*.

Collis, B. A., Knezek, G. A., Lai, K., Miyashita, K. T., Pelgrum, W. J., Plomp, T., & Sakamoto, T. (1996). *Children and Computers in School.* Mahwah, NJ: Erlbaum.

*Education Week* (January 1999). "Quality Counts '99."

Elley, W. B. (1993). *International Report: The IEA Study of Literature: Achievement and Instruction in Thirty-Two School Systems.* Oxford: Pergamon.

Ercikan, K. (1997). "Linking Statewide Tests to the National Assessment of Educational Progress: Accuracy of Combining Test Results Across States." *Applied Measurement in Education, 10,* 145–159.

Gearhart, M., Herman, J. L., Baker, E. L., & Whittaker, A. K. (1994). *Whose Work Is It? A Question for the Validity of Large-Scale Portfolio Assessment.* CSE Technical Report 363. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Hambleton, R. K., Jaeger, R. M., Koretz, D., Linn, R. L., Millman, J., & Phillips, S. E. (1995, June). *Review of the Measurement Quality of the Kentucky Instructional Results Information System, 1991–1994.* Report prepared for the Office of Educational Accountability, Kentucky General Assembly.

Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). "Interview Procedures for Validating Science Assessments." *Applied Measurement in Education, 10,* 181–200.

Jones, L. (1997). "National Tests and Education Reform: Are They Compatible?" ETS William H. Angoff Memorial Lecture. Princeton, NJ: Educational Testing Service.

Klein, S. P., McCaffrey, D. M., Stecher, B., & Koretz, D. (1995). "The Reliability of Mathematics Portfolio Scores: Lessons from the Vermont Experience." *Applied Measurement in Education, 8,* 243–260.

Kolen, M. J., & Brennan, R. L. (1995). *Test Equating.* New York: Springer.

Koretz, D., and Barron, S. I. (1998). *The Validity of Gains on the Kentucky Instructional Results Information System (KIRIS).* Santa Monica, CA: RAND.

Koretz, D., Linn, R. L., Dunbar, S. B., and Shepard, L. A. (1991). "The Effects of High-Stakes Testing: Preliminary Evidence About Generalization Across Tests." In R. L. Linn (chair), *The Effects of High Stakes Testing.* Symposium presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago, April.

Koretz, D., Mitchell, K., Barron, S., & Keith, S. (1996). *Final Report: Perceived Effects of the Maryland School Performance Assessment Program.* CSE Technical Report 409. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). "The Vermont Portfolio Assessment Program: Findings and Implications." *Educational Measurement: Issues and Practice, 13* (3), 5–16.

Levine, R., Rathbun, A., Selden, R., & Davis, A. (1998). *NAEP's Constituents: What Do They Want? Report of the NAEP Constituents' Survey and Focus Groups.* NCES 98–521. Washington, DC: U.S. Department of Education.

Linn, R. L. (1993). "Linking Results of Distinct Assessments." *Applied Measurement in Education, 6*, 83–102.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). "Comparing State and District Test Results to National Norms: The Validity of Claims That 'Everyone Is Above Average.'" *Educational Measurement: Issues and Practice, 9*, 5–14.

Linn, R. L., & Kiplinger, V. L. (1995). "Linking Statewide Tests to the National Assessment of Educational Progress: Stability of Results." *Applied Measurement in Education, 8*, 135–155.

Mead, A. D., & Drasgow, F. (1993). "Equivalence of Computerized and Paper-and-Pencil Cognitive Ability Tests: A Meta-analysis." *Psychological Bulletin, 114*, 449–458.

Mehrens, W. A., & Kaminski, J. (1989). "Methods for Improving Standardized Test Scores: Fruitful, Fruitless, or Fraudulent?" *Educational Measurement: Issues and Practice, 8* (1), 14–22.

Mullis, I.V.F., Martin, M. O., Beaton, A. E., Gonzales, E. J., Kelly, D. L., & Smith, T. A. (1998). *Mathematics and Science Achievement in the Final Year of Secondary School: IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA: TIMSS International Study Center.

National Center for Education Statistics (1998). *Internet Access in Public Schools.* Washington, DC: U.S. Department of Education.

National Council on Education Standards and Testing (1992). *Raising Standards for American Education.* Washington, DC: U.S. Government Printing Office.

National Research Council, Committee on Appropriate Test Use (1998). *High Stakes: Tests for Tracking, Promotion, and Graduation.* Washington, DC: National Academy Press.

National Research Council, Committee on Equivalency and Linkage of Educational Tests (1998). *Uncommon Measures: Equivalence and Linkage Among Educational Tests.* Washington, DC: National Academy Press.

National Research Council and National Council of Teachers of Mathematics (1997). *Improving Student Learning in Mathematics and Science: The Role of National Standards in State Policy.* Washington, DC: National Academy Press.

O'Neil, H. F., Sugrue, B., Abedi, J., Baker, E. L., & Golan, S. (1997). *Final Report on Experimental Studies of Motivation and NAEP Test Performance.* CSE Technical Report 427. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Owings, J. (1995). *Psychometric Report for the NELS:88 Base Year Through Second Follow-Up.* NCES 95382. Washington, DC: National Center for Education Statistics.

Pashley, P. J., & Phillips, G. W. (1993). *Toward World-Class Standards: A Research Study Linking International and National Assessments.* Princeton, NJ: Educational Testing Service.

Russell, M., & Haney, W. (1997). "Testing Writing on Computers: An Experiment Comparing Student Performance on Tests Conducted via Computer and via Paper-and-Pencil." *Education Policy Analysis Archives, 5* (3); available at http://olam.ed.asu.edu/epaa/.

Smith, M. L. (1991). "Meanings of Test Preparation." *American Educational Research Journal, 38*, 521–542.

Smith, M. S., Stevenson, D. L., & Li, C. P. (1998). "Voluntary National Tests Would Improve Education." *Educational Leadership, 55* (6), 42–44.

Stake, R. E. (1995). "The Invalidity of Standardized Testing for Measuring Mathematics Achievement." In Thomas A. Romberg (ed.), *Reform in School Mathematics and Authentic Assessment* (pp. 173–235). Albany, NY: SUNY Press.

Stecher, B. M., Barron, S., Kaganoff, T., & Goodwin, J. (1998). *The Effects of Standards-Based Assessment on Classroom Practices: Results of the 1996–97 RAND Survey of Kentucky Teachers of Mathematics and Writing.* CSE Technical Report 482. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Stecher, B. M., & Klein, S. P. (1997). "The Cost of Science Performance Assessments in Large-Scale Testing Programs." *Educational Evaluation and Policy Analysis, 19*, 1–14.

Stecklow, S. (1997, September 2). "Kentucky's Teachers Get Bonuses, but Some Are Caught Cheating." *The Wall Street Journal*, pp. A1, A5.

U. S. Department of Education (1997). *The NAEP Guide: A Description of the Content and Methods of the 1997 and 1998 Assessments.* NCES 97990. Washington, DC: National Center for Education Statistics.

U. S. Department of Education and National Science Foundation (1998). *Action Strategy for Improving Achievement in Mathematics and Science.*

Wainer, H. (1993). "Measurement Problems." *Journal of Educational Measurement, 30,* 1–21.

Wainer, H., & Thissen, D. (1993). "Combining Multiple-Choice and Constructed-Response Test Scores: Toward a Marxist Theory of Test Construction." *Applied Measurement in Education, 6,* 103–118.

Waltman, K. K. (1997). "Using Performance Standards to Link Statewide Achievement Results to NAEP." *Journal of Educational Measurement, 34,* 101–121.

Wenglinsky, H. (1998). *Does It Compute? The Relationship Between Educational Technology and Student Achievement in Mathematics.* ETS Policy Information Report. Princeton, NJ: Educational Testing Service.

Wolf, L. F., & Smith, J. K. (1995). "The Consequence of Consequence: Motivation, Anxiety, and Test Performance." *Applied Measurement in Education, 8,* 227–242.

RAND Education
1700 Main Street
P.O. Box 2138
Santa Monica, CA 90407-2138
310/393-0411